Characterization of PET/CT images using texture analysis: the past, the present... any future?

Mathieu Hatt, Florent Tixier, Larry Pierce, Paul E Kinahan, Catherine Cheze Le Rest, Dimitris Visvikis

▶ **To cite this version:**

Mathieu Hatt, Florent Tixier, Larry Pierce, Paul E Kinahan, Catherine Cheze Le Rest, et al.. Characterization of PET/CT images using texture analysis: the past, the present... any future?. European Journal of Nuclear Medicine and Molecular Imaging, 2016, 10.1007/s00259-016-3427-0 . hal-01330349

HAL Id: hal-01330349
https://hal.science/hal-01330349

Submitted on 5 Oct 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Characterization of PET/CT images using texture analysis: the past, the present… any future?

Mathieu Hatt[1], PhD, Florent Tixier[2,3], PhD, Larry Pierce[4], PhD, Paul E. Kinahan[4], PhD, Catherine Cheze Le Rest[2,3], MD, PhD, Dimitris Visvikis[1], PhD.

[1] INSERM, UMR 1101, LaTIM, University of Brest, France.

[2] University Hospital, Nuclear Medicine, Poitiers, France.

[3] University of Poitiers, Medical school, EE DACTIM, Poitiers, France

[4] Imaging Research Laboratory, University of Washington, Seattle, USA.

Corresponding author:

Mathieu Hatt INSERM, UMR 1101, LaTIM

CHRU Morvan, 2 avenue Foch

29609, Brest, France

Tel: +33(0)2.98.01.81.11 - Fax: +33(0)2.98.01.81.24

e-mail: hatt@univ-brest.fr

Wordcount: ~10380 (including ~3800 for references and 230 for the abstract)

Abstract:

Introduction: After seminal papers in 2009-2011, the use of textural analysis of PET/CT images for quantification of intra-tumour uptake heterogeneity has gained interest in the last four years. Results are difficult to compare due to the heterogeneity of studies and lack of standardization. There are also numerous challenges to address.

Objective: The purpose of this review is to provide critical insights into the recent development of the use of texture analysis to quantify heterogeneity in PET/CT images, identify issues and challenges, as well as offer recommendations for their use in clinical research.

Discussion: Numerous potentially confounding issues have been identified, related to the complex workflow for textural features calculations, as well as the dependency of features on various factors such as acquisition, image reconstruction, pre-processing, functional volume segmentation, methods of establishing and quantifying correspondences with genomic and clinical metrics of interest. A lack of understanding of what the features may represent for underlying pathophysiological processes and the variability of technical implementation practices makes comparing results in the literature challenging, if not impossible. Since progress as a field requires pooling results, there is therefore an urgent need for standardization and recommendations/guidelines to enable the field to move forward. We provide a list of correct formulas for usual features and recommendations regarding implementation. Studies on larger cohorts with robust statistical analysis and machine learning approaches represent promising directions to evaluate the potential of this approach.

Keywords: PET/CT, image texture, heterogeneity, critical review, recommendations

Introduction

Tumours are heterogeneous entities at all scales (macroscopic, physiological, microscopic, genetic) [1]. As a multimodal imaging modality, Positron Emission Tomography/Computed Tomography (PET/CT) is a promising tool for noninvasive exploration of intra-tumour heterogeneity at a macroscopic scale in both the anatomical and functional dimensions [2,3]. The term heterogeneity usually conveys a different meaning depending on the image modality. When considering the PET component, it refers to the radiotracer uptake spatial distribution, which may reflect, depending on the radiotracer used, the combination of underlying biological processes such as metabolism, hypoxia, cellular proliferation, vascularization and necrosis [4–6]. Regarding the low-dose, usually non-contrasted, CT component of PET/CT, heterogeneity refers to the variability of tissue density, which may result from spatially varying vascularization, necrosis or cellularity, as well as the proportions of fat, air and water [7].

In other modalities such as contrast-enhanced CT, as well as in MRI various sequences (T1, T2, FLAIR, DCE-MRI…), heterogeneity can also include the spatial variability of vessel density, perfusion, proton density, physiological tissue characteristics, etc. [8–12].

The heterogeneity of image voxel intensities can be quantified by different image processing and analysis methods, including texture analysis (TA) [3], fractal analysis [13], shape models [14–16], intensity histogram analysis [15,17] or filtering combined with statistical and frequency-based methods [18]. This critical review will focus on the use of TA in PET/CT images, although for completeness a section dedicated to alternative heterogeneity metrics can be found in Supplemental material section 1.

Systematically constructing higher-dimensional information from data falls under the

general rubric of ‘-omics’, which includes genomics, proteomics and others [19]. Extracting a large amount of features from images (including TA metrics, shape descriptors and other quantitative metrics) has become popular under the denomination *radiomics* [20,21]. The potential of such an approach is to quantify properties of tissues and/or organs beyond the capability of visual interpretation or simple metrics. The use of TA has been widespread in MR and CT imaging since the early 90’s [22,23] and more recently (end of 2000’s) for PET intra-tumour heterogeneity characterization [15,24,25]. PET images have *a priori* less favorable properties for TA than MRI or CT, due to relatively lower signal-to-noise ratio and spatial resolution, as well as poorer spatial sampling. In addition, reconstructed PET images are often smoothed in clinical practice for the visual analysis by clinicians using filters which reduce the texture content of the image, such as the Gaussian filter [26]. However, the fidelity and quantitative accuracy of PET imaging has substantially improved in the last decade, with the advent of PET/CT systems, time-of-flight (TOF) capabilities, improved sensitivity, and the incorporation of several quantitative corrections in the current clinical gold standard iterative reconstruction algorithms [27]. Note also that the low-dose CT component of PET/CT has different image characteristics than the higher resolution dosimetry or diagnostic CT. In the last four years, dozens of studies investigating PET/CT uptake heterogeneity have been published and mostly focused on the PET component. Unfortunately, especially for TA, the number of required pre-processing steps and the numerous implementation choices in the involved workflow, have led to contradictory results and controversies, rendering impossible the comparison of results across studies [28]. In addition, the tendency in the biomedical field of publishing positive results easier than negative ones [29] may have led to overoptimistic interpretation of the potential value of TA in PET. This should also be

viewed in the context of current widespread concerns about reproducibility in biomedical research in general [30].

The objective of this review is to provide critical insights into the rapid development of the use of TA to characterize radiotracer uptake heterogeneity in PET/CT images, and identify current issues and challenges left to address. Finally, some recommendations for future research methodologies and reporting standards are discussed.

The past: Clinical potential and underlying motivations

The underlying motivation arose in part from the recognition that standard metrics considered in clinical practice or research studies, *i.e.* maximum or mean of standardized uptake value (SUV_{max} and SUV_{mean}) or the metabolically active tumour volume (MATV) do not fully describe the properties of tumours [14]. Some of these properties, such as shape and uptake heterogeneity, may reflect different tumour profiles associated with their aggressiveness, metastatic potential, or degree of response to a specific treatment, and consequently prognosis [31,32]. Quantifying these properties could provide indices with higher clinical value than usual metrics in stratifying patients or identifying poor responders to treatment. This proof-of-concept regarding the use of TA in PET images was shown first by El Naqa and colleagues in a seminal paper in 9 head and neck and 14 cervix cancer patients [15]. Only two other studies investigating TA in PET were published in the two following years. The first one demonstrated the impact of parameters used in PET iterative image reconstruction algorithms on TA metrics, of which many were shown to be sensitive to the resulting varying characteristics of the reconstructed images [24]. This highlighted the need for standardization if such features are to be considered within the context of multi-centric trials. Secondly, the extremely high variability (>100%) observed for some features suggested they should never be used, even in a single-site, single scanner study. The

second study reported on the predictive value of FDG uptake heterogeneity quantified using TA, in 41 patients with locally advanced esophageal cancer treated by concomitant chemoradiotherapy, showing that TA metrics had higher predictive value than SUV [25].

In subsequent works, several studies have shown significant correlations between the visual assessment of intra-tumour heterogeneity in PET images by experts and quantitative metrics including the area under the curve of the cumulative histogram [33], shape descriptors [34] (see Supplemental material section 1) and TA metrics [35]. However, the interpretation of the underlying biological meaning of PET image uptake heterogeneity and the explanation of why it may be potentially more powerful than other standard metrics is still largely based on assumptions linking it to differences in underlying metabolism, cellular proliferation, hypoxia and necrosis. This obviously depends on the radiotracer used, however the vast majority of the studies to date have been carried out using FDG (and static SUV images), with only a few examples on other radiotracers, such as FET [36], FLT [5,37], or DBTZ [16]. By comparison in CT or MRI, there have been several studies linking image-derived TA features with underlying pathophysiological properties including at the level of genomics [7,11,38–40], thereby providing growing evidence of their relevance and potential explanation for their observed clinical value. To the best of our knowledge, similar results currently available regarding TA applied to PET are very limited. A study established a correlation between perfusion CT derived parameters (*e.g.* blood flow) and TA metrics from FDG PET in stage III/IV colorectal tumours [41]. Regarding the relationship between PET TA features and data from underlying scales, preliminary results from a prospective study in 54 head and neck patients were recently presented, demonstrating that some PET TA metrics could be linked with altered signaling

pathways related for example to cell proliferation and apoptosis [42]. Such studies have the potential to help in understanding the observed higher clinical value of these metrics compared to standard quantitative parameters.

The present: An era of rapid expansion

Several dozens of papers investigating the clinical value of PET uptake heterogeneity (using TA or other methods) in various tumour types (including esophageal, lung, rectum, breast, head and neck, brain, lymphoma, etc.) as well as more recently in neurodegenerative diseases with PET [16,43] and DAT SPECT [44] have been published in the last four years alone. More recently, a few studies have also been interested in extracting features from both the PET and CT components (see the section “promising clinical results” below). For a more exhaustive list, we refer the readers to other recent reviews [2–4,9,45–49]. From a critical review of these studies, several common issues can be identified.

1. *Nomenclature variability, formula and implementation issues*

The variability in definitions of TA metrics and nomenclature, as well as errors in methods, published formulae and computational codes complicate any evaluation and comparison of the published results. The use of the term “textural” itself can be confusing: in a recent study, the title and abstract refer to “textural parameters”, reporting a higher predictive value regarding response to therapy in a cohort of 27 rectal cancer patients [50]. However, only 1st order histogram-derived features (coefficient of variation (COV), skewness and kurtosis) and none of the textural features of 2nd or higher order features (that actually take into account spatial distribution) were explored. In addition, since these metrics were compared in pre-, mid- and post-therapy PET images, their repeatability should be carefully verified. Yet

studies have reported before on a relatively low level of repeatability of these features, especially skewness [51,52]. This challenge is not restricted to PET studies, as a recent work on the use of DCE-MRI in lung cancer made the exact same use of the term “textural”, although only 1st order histogram-derived metrics were used [12]. Another example of potential nomenclature confusion is the use of the term “entropy” to mean “randomness” or “disorder” for the 1st order metric, when it is actually the entropy of the probability histogram [12]. Furthermore, we recommend to use the terms $entropy_{GLCM}$ and $entropy_{HIST}$ in order to avoid confusion between the feature calculated in the co-occurrence matrix and the one calculated in the histogram, as an intuitive understanding of entropy may not apply to these metrics. Similarly, we recommend to use the terms $constrast_{GLCM}$ and $constrat_{NGTDM}$ to avoid confusion.

Section 3 in the supplemental material contains a list of TA metrics calculation formula with detailed notes. Several software distributions have also been made available [53,54], but there is a need to ensure that all feature calculations are accurately implemented before they could be reliably used for research. Section 4 in the supplemental material contains a list of several such codes with associated remarks.

2. Workflow complexity

One issue with TA is the very large number of parameters that can theoretically be calculated, in some cases over 100, as well as the number of ways they can be calculated. The recognized sources of variability (acquisition protocol, scanner type, quantitative corrections, type of reconstruction algorithm and parameters, post-reconstruction image processing, region of interest definition, etc.) in the standard metrics (SUV, MATV) quantification can also have a similar impact on TA features. There are also additional steps and methodological choices that have a similarly (if not

higher) impact on the resulting TA metrics. Figure 1 illustrates the TA workflow complexity, with the different steps discussed in the following sections. Note than some upstream steps can also have an impact on these choices, such as the segmentation (*section 3 below*).

First order features estimate properties of individual voxel values, ignoring the spatial interaction between them (and as such cannot really be considered as “textural” features, because they do not differentiate spatial arrangements and patterns), whereas second- and higher-order features estimate properties of two or more voxel values occurring at specific locations relative to each other. For these 2^{nd} and higher order TA features, the first steps usually consist in re-sampling or interpolating the non-cubic voxel grid into cubic voxels (seldom carried out) and performing quantization (systematically carried out, also called *discretization*, *downsampling* or *resampling*) of the original intensities (or SUV) into a discrete set of values. This number determines the size of the matrices that are built and in which TA metrics are subsequently calculated. Several methods have been proposed to perform this quantization (see supplemental section 2) such as a linear distribution into a set number of bins (*e.g.* 32 or 64) [15,25], the use of a clustering algorithm (Max-Lloyd) [55] or into bins of fixed width (*e.g.* 0.25 [52] or 0.5 [56,57] SUV). The chosen quantization approach and value can have an important impact on the resulting TA metrics, as well as their relationship with tumor volume or SUV_{max} [51,56,58–60] and it is thus an important factor not to overlook, as illustrated in figure 2.

The second step consists in building texture matrices, of which several exist (e.g. grey-level co-occurrence matrix GLCM, neighborhood grey tone difference matrix NGTDM and grey level zone size matrix GLZSM) and can be built in different ways (see supplemental material, section 3). For example, co-occurrence matrices quantify

relationships between pairs of voxels. They are usually defined according to a given spatial direction and a given distance between the pairs of voxels. For a 3D analysis, 13 directions are often considered and one matrix is built per direction. The TA metric is then calculated in each of these matrices, and the 13 resulting values are averaged. Usually, the distance is set to one voxel. Modifying these choices (*e.g.* using only one matrix for all directions) can lead to different TA features distributions (see figure 3), associated complementary value with other metrics, and as a consequence, overall clinical value [55,60].

3. PET tumour volume segmentation

Numerous studies have used the least robust and/or accurate methods to define overall tumour volume, such as manual delineation or fixed thresholding. Single observer manual delineation suffers from high inter- and intra-observer variability, whereas fixed thresholding significantly underestimates the true MATV extent by focusing on the tumour sub-volume with the highest uptake [61,62]. This in turn may bias the heterogeneity assessment and the associated ranking of intra-tumour heterogeneity levels. Another issue concerns the way the tumour volume is *a priori* considered in the analysis. More specifically one can define functional volume so that areas with low or no radiotracer uptake are included in the volume, or alternatively excluded from it. Excluding these areas would exclude necrotic regions but would also limit the risk of including non-pathological areas in the heterogeneity analysis. The choice of the segmentation approach used may result in more or less constraints. For example, with a gradient-based tool [63,64], the resulting contour is binary only and covers the entire tumour including areas without uptake (figure 4). On the contrary, with a method based on region-growing or clustering paradigms [65,66] the areas with

uptake similar to the background are usually excluded, although they could easily be included in the analysis with an additional step.

4. Statistical issues

In the vast majority of the published studies, there is no multivariate analysis including potential confounding factors, nor a correction for multiple testing, and very rarely was robust machine learning with cross-validation used. The size of the patient cohorts considered is most often very small with respect to the number of explored parameters and tested hypotheses. Ideally, these studies should be combined to provide a meta-analysis; however, there are often problems with how results are reported [21] which renders such a meta-analysis practically impossible. Finally, the cohorts are often heterogeneous in terms of staging or treatment modality, are most often retrospective in nature, and the results are almost never validated on external cohorts. A recent review paper highlighted these issues for a selection of 15 studies (in both PET and CT), showing that the majority of these suffered from at least some of these shortcomings. The review concluded that the clinical value of TA metrics extracted from CT or PET images remains to be demonstrated [28]. Although the bias in publishing positive results is strong in the biomedical field [29], one has to keep in mind that only a handful of studies have concluded that heterogeneity quantification does not bring any value regarding the aimed clinical endpoint [67–69], with the overall trend being mostly positive. More specifically, two studies in cervical cancer have concluded that a metric based on a “Volume versus Threshold Curve” was not able to predict outcome in 73 patients [67], contradicting a previous assessment in the same cohort [70]. Although other 1st order features such as standard deviation, skewness and kurtosis were also included, it should be emphasized that this study essentially highlighted the fact that the metric based on “volume versus threshold curve” is a

surrogate of volume, not a measurement of heterogeneity (see also supplemental material). The same authors further explored additional metrics (sphericity, extent, Shannon entropy and the accrued deviation from smoothest gradients, *i.e.* not TA metrics) in another group of 85 FIGO stage IIb patients, with similar negative conclusions regarding the prediction of pelvic lymph node involvement [69]. Finally, in contradiction to these two studies, the same group also published another paper in which TA metrics had predictive value of response to therapy in 20 cervical cancer patients when considering their temporal evolution between baseline, week 2, week 4 and post-therapy PET scans [71]. It should also be noted that in all these studies MATV were delineated using a fixed threshold at 40% of SUV_{max}.

A recent study in breast cancer showed in a prospective homogeneous cohort of 171 women, that contrary to previous results obtained in a smaller cohort (n=54) [72], none of the considered PET TA metrics were able to improve differentiation between the three main molecular subtypes of breast tumours beyond the standard clinical factors and SUV metrics [68].

5. Redundancy of features

The vast majority of studies are based on analyzing a pre-determined functional tumour volume, which is thus known prior to the heterogeneity characterization. Therefore an heterogeneity metric can only have complementary (or significantly higher) value if it is not highly correlated with the corresponding volume. The correlation between heterogeneity metrics and the MATV or another image derived parameter (such as SUV_{max}, SUV_{mean} or Total Lesion Glycolysis=$MATV \times SUV_{mean}$ [73]) can be explained by two different but complementary aspects: the mathematical/algorithmic design of the parameter, and the fact that heterogeneity is intrinsically and biologically correlated

with it. Indeed, heterogeneity quantification of uptake is expected to be correlated with other tumour properties. In most solid tumours larger volumes exhibit a wider range of heterogeneity patterns and intensity than smaller ones. This is due first to the fact that larger tumours have more potential to be composed of several different types of tissues and regions with variable uptake that can be resolved in a PET image compared to smaller volumes, for which a similar heterogeneity may exist at the cellular and tissue levels but is blurred due to the limited spatial resolution. On the other hand, a correlation between high heterogeneity and high SUV seems less logical, since small homogeneous uptakes can have high or low SUV_{max}, whereas both larger homogeneous or heterogeneous lesions can exhibit a wide range of maximum uptake. The challenge for future studies is therefore to identify which part of the correlation comes from a biological reality, imaging limitations and/or from the mathematical and algorithmic definition of the heterogeneity metrics. Another challenge is to identify above what level of correlation to exclude a TA feature for a subsequent multivariate analysis since it is unlikely to provide complementary information. This is less trivial than it may sound, since the absolute level of correlation varies depending on the chosen coefficients: Pearson coefficients may significantly underestimate the correlation between two metrics when it is not linear. Kendall and Spearman coefficients both provide rank correlation assessment, however Kendall coefficients are usually smaller than Spearman, as illustrated in figure 5. These would thus require different scales to describe strong, moderate or weak correlations (in the case presented in figure 5, the correlation is above 0.9 according to Spearman but is below 0.8 according to Kendall). Using simple correlation coefficients to select features to combine in a multi-parametric model may be sub-optimal, and we recommend using

robust machine learning techniques to achieve better redundancy analysis and feature selection / combination.

Although it should be recognized that the relationship between tumour volume and PET spatial resolution has an impact on derived metrics, there is an additional factor to take into account for TA metrics that quantify intensity and spatial relationships between pairs or groups of voxels. The correlation between a TA metric and the volume of interest in which it is calculated needs to be analyzed not only in terms of absolute volume, but more importantly in terms of the number of voxels involved in the calculation. For example, if we consider a given tumour volume sampled on a 2×2×2 mm^3 or 4×4×4 mm^3 grid, the $entropy_{GLCM}$ metric will have higher value for the 2×2×2 mm^3 image compared to the 4×4×4 mm^3 one, not because of a higher heterogeneity, but only because of a higher number of voxels involved in the calculation. This “number of voxels confounding effect” has been demonstrated in a recent study that showed the matrix grid in the reconstruction has a strong impact on most TA metrics [74]. It is especially crucial to take into account in a multi-centric study where reconstructed matrix sizes vary across sites. Regarding the tumour volume confounding effect, the correlation between TA metrics and tumour volume was reported in several studies [59,60,75,76], two of which were specifically focused on the issue. The first aimed to identify which was the minimal volume of interest so that the volume would be sufficiently large for the TA to differentiate between different levels of heterogeneity and concluded that a minimal volume of 45 cm^3 was necessary [76]. This study was based on the use of a single TA feature, with a uniform quantization (Q=152 bins), an analysis in 2D with 2 directions (horizontal and vertical), with one co-occurrence matrix used for each direction followed by averaging the resulting values, and without testing different configurations. Another study tackled the same issue by considering 555

tumours of 5 different types [60]. The results showed that the correlation between TA metrics and volume was highly variable among TA features, decreased as tumour volumes increased, and depended on the quantization value and design of the co-occurrence matrices. The above study also showed that using the same set of parameters as in the previous study [76] led to a very high correlation (>0.9) between the volume and $entropy_{GLCM}$, whereas using a different configuration (a smaller quantization value of 64 and only one co-occurrence matrix taking into consideration all directions simultaneously instead of averaging the values obtained in different matrices) the correlation dropped below 0.5. With this calculation, $entropy_{GLCM}$ thus provided complementary information to volume (figure 3), which also translated into complementary prognostic value: when the two parameters were combined, improved stratification of patients overall survival could be reached in 101 NSCLC patients [60]. These two studies clearly highlight the need to carefully consider interactions between tumour volume and heterogeneity quantitative metrics. The latter study also showed that the choices made in the calculation workflow can determine the efficacy of a metric.

Beyond the relationships with tumour volume, TA provides the ability to calculate numerous parameters that also exhibit high levels of redundancy [59,75]. It is therefore necessary to establish a method of selection amongst all calculated features (and amongst all ways of calculating them). The properties of an “ideal” heterogeneity metric as well as the recommended methodology to assess them are listed in table 1.

6. Repeatability and reproducibility / robustness

Repeatability is a measurement of precision that occurs with identical or near-identical conditions, *e.g.* double baseline (also called test-retest) studies. Reproducibility, in

contrast, is a measurement of precision when location, measuring system, or other factors differ [77]. In reproducibility studies, the objective is to measure the effects of different conditions on the performance of a quantitative imaging biomarker with the goal of demonstrating equivalent performance in less restrictive study conditions.

Repeatability has already been studied for TA in PET, demonstrating that only a handful of metrics have repeatability limits similar to MATV and SUV measurements, and are therefore reliable enough to be considered further, especially in the context of therapy monitoring [51,52,57]. Regarding reproducibility or robustness, it has been shown that only a few features are robust versus variations in image reconstruction algorithm types [24,52,74,78]. It was also shown that the impact of tumour segmentation [52,57,58,75], post-reconstruction smoothing [26], quantization [52,58] or partial volume effects correction [75] varied between TA metrics. More recently, two studies also investigated the impact of respiratory motion, by comparing TA features in PET images of lung cancer patients with and without respiratory gating, showing that TA features may be affected in standard non-gated acquisitions, particularly in the lower lung lobes [79,80]. It was also shown that the use of a respiration-averaged CT instead of a helical CT for attenuation correction of the PET data had a higher impact on SUV and TLG than on TA metrics [81]. Another study demonstrated that TA features calculated in parametric maps derived from dynamic PET acquisitions or from corresponding static SUV images were not significantly different, suggesting that heterogeneity quantification on parametric images through TA may not provide significant additional information compared to static SUV images [82]. Finally, it was also shown that even basic stochastic effects of PET acquisitions can affect some TA metrics [83]. All these factors may impact not only the absolute values of calculated features, but also their correlation with volume. Figure 6 illustrates this on a set of

tumours reconstructed using two different voxel sizes (the only modified parameter in the reconstruction). The exact same number of voxels was considered using either nearest neighbors or b-splines interpolation of the image with larger voxels, in order to obtain exactly the same number of voxels as the image with the smaller voxels.

Promising clinical results

Demonstration of the clinical value of TA requires large cohorts of patients combined with a rigorous statistical analysis. Although a number of studies have been recently published exploiting cohorts between 20 and 70 patients only [50,71,84–90], some of the most recent studies have been carried out on larger cohorts between 80 and more than 200 patients: 88 oropharyngeal squamous cell carcinoma [91], 103 bone and soft tissue lesions [92], 101 early-stage NSCLC [93], 112 œsophageal and 101 NSCLC [60], 113 gliomas [36], 107 and 217 œsophageal [94,95], 132 lymph nodes in lung cancer [96], 116, 195 and 201 NSCLC [97–99], 137 pancreatic lesions [100] and 188 lesions in lymphoma patients [101]. Some of the most recent studies have also used more robust statistical analysis, compared to these recently reviewed [28], several of them using machine learning method, *e.g.* neural networks [94], support vector machines [92,96,101] or the least absolute shrinkage and selection operator (LASSO) [93,99]. The majority of these recent studies have reached the conclusion that TA can provide useful quantitative metrics regarding patient management (prognosis, response to therapy, distant metastasis prediction…) in different cancer models except one that showed more mixed results [102], whereas another concluded that the improvement, although significant, may not be sufficient to have a clinical impact [95].

Note also that a few studies have recently investigated the potential combination of image-derived features from both PET and the corresponding low-dose CT component

of the PET/CT dataset [8,92,96,97,101,103], while another recent work has combined TA from PET and MRI to predict lung metastases in soft-tissue sarcomas of the extremities [55]. A recent proof-of-concept study (on 2 patients) investigated the characterization of renal cell carcinoma in simultaneous ^{18}F-FLT PET/MRI acquisitions (with pre- and mid-treatment images) [104].

Finally, the combination of PET image-derived features with other contextual data may be considered, such as in a recent study in which zone-size non-uniformity extracted from pretreatment FDG PET was combined with key immunohistochemistry metrics, leading to improved stratification in 113 patients with advanced stage oropharyngeal squamous cell carcinoma [105]. In another recent work, TA metrics extracted from both PET and low-dose CT components of PET/CT standard acquisitions were combined to improve patient prognosis stratification over clinical staging [97].

When dealing with a large number of variables and limited size cohorts, robust methods for features selection combined with an appropriate classifier and testing with cross-validation can provide tools with good performance. One has to note however that validation using an external cohort remains the gold standard, although it is still rarely performed in PET/CT studies. In a recent work, 31 patients out of 101 were used for validation, while the first 70 were used for building the model [93]. Contrary to the PET/CT field, the use of machine learning techniques including robust features selection, combination of features within classifiers and cross-validation or validation in an external cohort is well established in the fields of CT [7,106–108] and MR [109–113] radiomics. To date, only a limited number of clinical studies investigating TA in PET/CT have exploited up-to-date techniques from the field of machine learning. Recently, a more technical study showed that appropriate hierarchical forward selection of features combined with a support vector machine classifier could improve

results [114]. A recent study focused on this topic and compared 14 feature selection methods and 12 classifiers using large cohorts of CT datasets, ranking the relative performance in terms of accuracy and stability of each approach [107]. An interesting result is that the performance was mostly affected by the choice of the classifier (34% of total variance). Although not yet performed on PET datasets, this work could nonetheless form a basis for selecting appropriate machine learning methods for future studies investigating the value of TA in PET imaging.

Conclusions: a future for texture analysis in PET?

Although not impossible, it is challenging to compare results from the numerous studies currently available and draw concrete conclusions on the clinical value of TA in PET imaging. This is due to a large variability in the implemented methodology associated with the workflow complexity involved with the calculation of features, in combination with the lack of technical details provided in most studies. In addition, numerous issues related to the statistical analysis and bias in publishing mostly positive results further increases the difficulty in drawing firm conclusions from the currently published literature. Keeping in mind these current limitations, it should nonetheless be emphasized that most of the studies currently published point to complementary and additional value in extracting more advanced image features from medical images, including PET/CT. Although the level of evidence is likely insufficient today, a positive trend can be observed. Thus the use of TA in PET/CT images should not be abandoned but rather reinforced by increasing the level of requirement for the studies to be conducted, in order to enable the field to move forward. Table 2 lists several objectives we should consider as a community to achieve this.

Establishing a benchmark (objective 1) could start by providing users with tools to validate features calculation codes. Test images with known and verified associated metrics values for a range of calculation methods and choices could be provided, so that users can validate their TA metrics implementations [83].

The next step (objective 2) should consist in generating and circulating a draft of recommendations and guidelines regarding choices in pre-processing and segmentation steps, especially within the context of multicentric studies. Although variability inherent in merging results based on images from various devices and reconstructions algorithms from different vendors may be difficult to avoid, some recommendations can already be made based on current results. First, pre-process and resample images to a common voxel size, preferably isotropic [55]. This would allow avoiding major issues in comparing co-occurrence derived metrics calculated on different spatial sampling [55,74]. Second, avoid post-reconstruction smoothing altogether, or use appropriate edge-preserving filters [26]. Third, automated segmentation approaches robust to heterogeneous distribution should be used [61,66]. Alternatively, if only fixed or adaptive thresholding methods are available, manual / visual checking and editing should be mandatory to avoid under-segmentation of heterogeneous uptakes.

Objective 3 could consist in establishing recommendations on which features should be preferably used and those to exclude (*e.g.* features that have been identified with very poor repeatability and robustness), as well as recommended workflow choices to obtain features with lowest redundancy and highest clinical value. In that regard we provided in the supplemental material a list of verified formulas for the usual features, as well as comments, corrections and implementation recommendations to avoid

common mistakes and misconceptions. We also recommend to preferably rely on features that have been demonstrated as robust and repeatable [51,52,57,58,75].

Finally, beyond providing test images, objects, and open-source codes and formulae to improve standardization between research groups (objective 1), it would be beneficial that a benchmark would also contain publicly available clinical datasets of PET/CT images along with clinical endpoint information (prognosis, response to therapy, tumour type, etc.) and other clinical data (objective 4). This would allow any research group to test its own workflow (image pre-processing, tumour segmentation, TA metrics calculation, machine learning features selection and classifiers, etc.) and then compare its results with those by other groups. The Cancer Imaging Archive (TCIA) could be the support of such future efforts, as it already contains several publicly available cohorts of patients with images of various modalities and the associated clinical data.

Further efforts along these lines could be organized within taskgroups of the EANM, QIBA, QIN, SNMMI and AAPM and they should happen in the nearest future if TA (and *radiomics* in general) are to have any future in PET/CT imaging.

Acknowledgements

The authors wish to thank Issam El Naqa, Philippe Lambin, Hugo Aerts, Ralph Leijenaar, Floris Van Velden, Martin Vallières, Arman Rahmim, Matt Nyflot, Art Chaovalitwongse, as well as the members of the RSNA Quantitative Imaging Biomarkers Alliance and the NCI Quantitative Imaging Network for many helpful discussions.

Supported in part by NIH grant U01-CA148131.

Compliance with Ethical Standards

Funding: This work has received a French government support granted to the CominLabs excellence laboratory and managed by the National Research Agency in the "Investing for the Future" program under reference ANR-10-LABX-07-01.

Disclosure of Conflicts of Interest: The authors declare that they have no conflict of interest.

PEK has a research grant from GE Healthcare and is co-founder of PET/X LLC.

Research involving Human Participants and/or Animals: not applicable.

References

1. Gerlinger M, Rowan AJ, Horswell S, Larkin J, Endesfelder D, Gronroos E, et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. N Engl J Med. 2012;366:883–92.

2. Davnall F, Yip CS, Ljungqvist G, Selmi M, Ng F, Sanghera B, et al. Assessment of tumor heterogeneity: an emerging imaging tool for clinical practice? Insights Imaging. 2012;3:573–89.

3. Chicklore S, Goh V, Siddique M, Roy A, Marsden PK, Cook GJ. Quantifying tumour heterogeneity in 18F-FDG PET/CT imaging by texture analysis. Eur J Nucl Med Mol Imaging. 2013;40:133–40.

4. O'Connor JPB, Rose CJ, Waterton JC, Carano RAD, Parker GJM, Jackson A. Imaging Intratumor Heterogeneity: Role in Therapy Response, Resistance, and Clinical Outcome. Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res. 2015;21:249–57.

5. Willaime JMY, Turkheimer FE, Kenny LM, Aboagye EO. Quantification of intra-tumour cell proliferation heterogeneity using imaging descriptors of 18F fluorothymidine-positron emission tomography. Phys. Med. Biol. 2013;58:187.

6. Weber WA, Schwaiger M, Avril N. Quantitative assessment of tumor metabolism using FDG-PET imaging. Nucl. Med. Biol. 2000;27:683–7.

7. Aerts HJWL, Velazquez ER, Leijenaar RTH, Parmar C, Grossmann P, Cavalho S, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat. Commun. 2014;5:4006.

8. Win T, Miles KA, Janes SM, Ganeshan B, Shastry M, Endozo R, et al. Tumor Heterogeneity and Permeability as Measured on the CT Component of PET/CT Predict Survival in Patients with Non-Small Cell Lung Cancer. Clin. Cancer Res. 2013;19:3591–9.

9. Asselin M-C, O'Connor JPB, Boellaard R, Thacker NA, Jackson A. Quantifying heterogeneity in human tumours using MRI and PET. Eur. J. Cancer Oxf. Engl. 1990. 2012;48:447–55.

10. O'Connor JP, Rose CJ, Jackson A, Watson Y, Cheung S, Maders F, et al. DCE-MRI biomarkers of tumour heterogeneity predict CRC liver metastasis shrinkage following bevacizumab and FOLFOX-6. Br J Cancer. 2011;105:139–45.

11. Nicolasjilwan M, Hu Y, Yan C, Meerzaman D, Holder CA, Gutman D, et al. Addition of MR imaging features and genetic biomarkers strengthens glioblastoma survival prediction in TCGA patients. J. Neuroradiol. 2015;42:212–21.

12. Yoon SH, Park CM, Park SJ, Yoon J-H, Hahn S, Goo JM. Tumor Heterogeneity in Lung Cancer: Assessment with Dynamic Contrast-enhanced MR Imaging. Radiology. 2016;151367.

13. Michallek F, Dewey M. Fractal analysis in radiological and nuclear medicine perfusion imaging: a systematic review. Eur. Radiol. 2014;24:60–9.

14. O'Sullivan F, Roy S, Eary J. A statistical measure of tissue heterogeneity with application to 3D PET sarcoma data. Biostatistics. 2003;4:433–48.

15. El Naqa I, Grigsby P, Apte A, Kidd E, Donnelly E, Khullar D, et al. Exploring feature-based approaches in PET images for predicting cancer treatment outcomes. Pattern Recognit. 2009;42:1162–71.

16. Gonzalez ME, Dinelle K, Vafai N, Heffernan N, McKenzie J, Appel-Cresswell S, et al. Novel spatial analysis method for PET images using 3D moment invariants: applications to Parkinson's disease. NeuroImage. 2013;68:11–21.

17. van Velden FH, Cheebsumon P, Yaqub M, Smit EF, Hoekstra OS, Lammertsma AA, et al. Evaluation of a cumulative SUV-volume histogram method for parameterizing heterogeneous intratumoural FDG uptake in non-small cell lung cancer PET studies. Eur J Nucl Med Mol Imaging. 2011;38:1636–47.

18. Ganeshan B, Miles KA. Quantifying tumour heterogeneity with CT. Cancer Imaging Off. Publ. Int. Cancer Imaging Soc. 2013;13:140–9.

19. Committee on the Review of Omics-Based Tests for Predicting Patient Outcomes in Clinical Trials, Board on Health Care Services, Board on Health Sciences Policy, Institute of Medicine. Evolution of Translational Omics: Lessons Learned and the Path Forward [Internet]. Micheel CM, Nass SJ, Omenn GS, editors. Washington (DC): National Academies Press (US); 2012 [cited 2016 Feb 23]. Available from: http://www.ncbi.nlm.nih.gov/books/NBK202168/

20. Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RG, Granton P, et al. Radiomics: extracting more information from medical images using advanced feature analysis. Eur J Cancer. 2012;48:441–6.

21. Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images Are More than Pictures, They Are Data. Radiology. 2016;278:563–77.

22. Mir AH, Hanmandlu M, Tandon SN. Texture analysis of CT-images for early detection of liver malignancy. Biomed. Sci. Instrum. 1995;31:213–7.

23. Schad LR, Blüml S, Zuna I. MR tissue characterization of intracranial tumors by means of texture analysis. Magn. Reson. Imaging. 1993;11:889–96.

24. Galavis PE, Hollensen C, Jallow N, Paliwal B, Jeraj R. Variability of textural features in FDG PET images due to different acquisition modes and reconstruction parameters. Acta Oncol. 2010;49:1012–6.

25. Tixier F, Le Rest CC, Hatt M, Albarghach N, Pradier O, Metges JP, et al. Intratumor heterogeneity characterized by textural features on baseline 18F-FDG PET images predicts response to concomitant radiochemotherapy in esophageal cancer. J Nucl Med. 2011;52:369–78.

26. Hatt M, Hanzouli H, Rest CCL, Visvikis D. Comparison of edge-preserving filters for unbiased quantification in 18F-FDG PET imaging. J. Nucl. Med. 2015;56:1828–1828.

27. Vaquero JJ, Kinahan P. Positron Emission Tomography: Current Challenges and Opportunities for Technological Advances in Clinical and Preclinical Imaging Systems. Annu. Rev. Biomed. Eng. 2015;17:385–414.

28. Chalkidou A, O'Doherty MJ, Marsden PK. False Discovery Rates in PET and CT Studies with Texture Features: A Systematic Review. PloS One. 2015;10:e0124165.

29. Dwan K, Gamble C, Williamson PR, Kirkham JJ, Reporting Bias Group. Systematic review of the empirical evidence of study publication bias and outcome reporting bias - an updated review. PloS One. 2013;8:e66844.

30. Ioannidis JPA. How to make more published research true. PLoS Med. 2014;11:e1001747.

31. Basu S, Kwee TC, Gatenby R, Saboury B, Torigian DA, Alavi A. Evolving role of molecular imaging with PET in detecting and characterizing heterogeneity of cancer tissue at the primary and metastatic sites, a plausible explanation for failed attempts to cure malignant disorders. Eur J Nucl Med Mol Imaging. 2011;38:987–91.

32. Visvikis D, Hatt M, Tixier F, Cheze Le Rest C. The age of reason for FDG PET image-derived indices. Eur J Nucl Med Mol Imaging. 2012;39:1670–2.

33. Watabe T, Tatsumi M, Watabe H, Isohashi K, Kato H, Yanagawa M, et al. Intratumoral heterogeneity of F-18 FDG uptake differentiates between gastrointestinal stromal tumors and abdominal malignant lymphomas on PET/CT. Ann. Nucl. Med. 2012;26:222–7.

34. Kim D-H, Jung J-H, Son SH, Kim C-Y, Jeong SY, Lee S-W, et al. Quantification of Intratumoral Metabolic Macroheterogeneity on 18F-FDG PET/CT and Its Prognostic Significance in Pathologic N0 Squamous Cell Lung Carcinoma. Clin. Nucl. Med. 2015;

35. Tixier F, Hatt M, Valla C, Fleury V, Lamour C, Ezzouhri S, et al. Visual versus quantitative assessment of intratumor 18F-FDG PET uptake heterogeneity: prognostic value in non-small cell lung cancer. J. Nucl. Med. 2014;55:1235–41.

36. Pyka T, Gempt J, Hiob D, Ringel F, Schlegel J, Bette S, et al. Textural analysis of pre-therapeutic [18F]-FET-PET and its correlation with tumor grade and patient survival in high-grade gliomas. Eur. J. Nucl. Med. Mol. Imaging. 2015;

37. Majdoub M, Visvikis D, Tixier F, Hoeben B, Visser E, Cheze Le Rest C, et al. Proliferative 18F-FLT PET tumor volumes characterization for prediction of locoregional recurrence and disease-free survival in head and neck cancer. Soc. Nucl. Med. Mol. Imaging Annu. Meet. 2013;

38. Segal E, Sirlin CB, Ooi C, Adler AS, Gollub J, Chen X, et al. Decoding global gene expression programs in liver cancer by noninvasive imaging. Nat Biotechnol. 2007;25:675–80.

39. Gevaert O, Mitchell LA, Achrol AS, Xu J, Echegaray S, Steinberg GK, et al. Glioblastoma multiforme: exploratory radiogenomic analysis by using quantitative image features. Radiology. 2014;273:168–74.

40. Wan T, Bloch BN, Plecha D, Thompson CL, Gilmore H, Jaffe C, et al. A Radio-genomics Approach for Identifying High Risk Estrogen Receptor-positive Breast Cancers on DCE-MRI: Preliminary Results in Predicting OncotypeDX Risk Scores. Sci. Rep. 2016;6:21394.

41. Tixier F, Groves AM, Goh V, Hatt M, Ingrand P, Le Rest CC, et al. Correlation of intra-tumor 18F-FDG uptake heterogeneity indices with perfusion CT derived parameters in colorectal cancer. PloS One. 2014;9:e99567.

42. Tixier F, Hatt M, Rest CCL, Simon B, Key S, Corcos L, et al. Signaling pathways alteration involved in head and neck cancer can be identified through textural features analysis in 18F-FDG PET images: a prospective study. J. Nucl. Med. 2015;56:449–449.

43. Klyuzhin IS, Gonzalez M, Shahinfard E, Vafai N, Sossi V. Exploring the use of shape and texture descriptors of positron emission tomography tracer distribution in imaging studies of neurodegenerative disease. J. Cereb. Blood Flow Metab. Off. J. Int. Soc. Cereb. Blood Flow Metab. 2015;

44. Rahmim A, Salimpour Y, Jain S, Blinder SAL, Klyuzhin IS, Smith GS, et al. Application of texture analysis to DAT SPECT imaging: Relationship to clinical assessments. NeuroImage Clin. [Internet]. [cited 2016 Apr 12]; Available from: http://www.sciencedirect.com/science/article/pii/S2213158216300341

45. Hatt M, Tixier F, Cheze Le Rest C, Visvikis D. Nouveaux indices en TEP/TDM : mythe et réalités. Médecine Nucl. 2015;39:331–8.

46. Carlier T, Bailly C. State-Of-The-Art and Recent Advances in Quantification for Therapeutic Follow-Up in Oncology Using PET. Front. Med. 2015;2:18.

47. Houshmand S, Salavati A, Hess S, Werner TJ, Alavi A, Zaidi H. An update on novel quantitative techniques in the context of evolving whole-body PET imaging. PET Clin. 2015;10:45–58.

48. Rahim MK, Kim SE, So H, Kim HJ, Cheon GJ, Lee ES, et al. Recent Trends in PET Image Interpretations Using Volumetric and Texture-based Quantification Methods in Nuclear Oncology. Nucl. Med. Mol. Imaging. 2014;48:1–15.

49. Cheng N-M, Fang Y-HD, Yen T-C. The promise and limits of PET texture analysis. Ann. Nucl. Med. 2013;27:867–9.

50. Bundschuh RA, Dinges J, Neumann L, Seyfried M, Zsótér N, Papp L, et al. Textural Parameters of Tumor Heterogeneity in ^{18}F-FDG PET/CT for Therapy Response Assessment and Prognosis in Patients with Locally Advanced Rectal Cancer. J. Nucl. Med. Off. Publ. Soc. Nucl. Med. 2014;55:891–7.

51. Tixier F, Hatt M, Le Rest CC, Le Pogam A, Corcos L, Visvikis D. Reproducibility of tumor uptake heterogeneity characterization through textural feature analysis in 18F-FDG PET. J Nucl Med. 2012;53:693–700.

52. van Velden FHP, Kramer GM, Frings V, Nissen IA, Mulder ER, de Langen AJ, et al. Repeatability of Radiomic Features in Non-Small-Cell Lung Cancer [(18)F]FDG-PET/CT Studies: Impact of Reconstruction and Delineation. Mol. Imaging Biol. MIB Off. Publ. Acad. Mol. Imaging. 2016;

53. Zhang L, Fried DV, Fave XJ, Hunter LA, Yang J, Court LE. IBEX: an open infrastructure software platform to facilitate collaborative work in radiomics. Med. Phys. 2015;42:1341–53.

54. Fang Y-HD, Lin C-Y, Shih M-J, Wang H-M, Ho T-Y, Liao C-T, et al. Development and evaluation of an open-source software package “CGITA” for quantifying tumor heterogeneity with molecular images. BioMed Res. Int. 2014;2014:248505.

55. Vallières M, Freeman CR, Skamene SR, El Naqa I. A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities. Phys. Med. Biol. 2015;60:5471–96.

56. Leijenaar RTH, Nalbantov G, Carvalho S, van Elmpt WJC, Troost EGC, Boellaard R, et al. The effect of SUV discretization in quantitative FDG-PET Radiomics: the need for standardized methodology in tumor texture analysis. Sci. Rep. 2015;5:11075.

57. Leijenaar RTH, Carvalho S, Velazquez ER, van Elmpt WJC, Parmar C, Hoekstra OS, et al. Stability of FDG-PET Radiomics features: an integrated analysis of test-retest and inter-observer variability. Acta Oncol. Stockh. Swed. 2013;52:1391–7.

58. Doumou G, Siddique M, Tsoumpas C, Goh V, Cook GJ. The precision of textural analysis in (18)F-FDG-PET scans of oesophageal cancer. Eur. Radiol. 2015;25:2805–12.

59. Orlhac F, Soussan M, Maisonobe J-A, Garcia CA, Vanderlinden B, Buvat I. Tumor Texture Analysis in 18F-FDG PET: Relationships Between Texture Parameters, Histogram Indices, Standardized Uptake Values, Metabolic Volumes, and Total Lesion Glycolysis. J Nucl Med. 2014;55:414–22.

60. Hatt M, Majdoub M, Vallières M, Tixier F, Le Rest CC, Groheux D, et al. 18F-FDG PET Uptake Characterization Through Texture Analysis: Investigating the Complementary Nature of Heterogeneity and Functional Tumor Volume in a Multi-Cancer Site Patient Cohort. J. Nucl. Med. 2015;56:38–44.

61. Hatt M, Cheze-le Rest C, van Baardwijk A, Lambin P, Pradier O, Visvikis D. Impact of tumor size and tracer uptake heterogeneity in (18)F-FDG PET and CT non-small cell lung cancer tumor delineation. J Nucl Med. 2011;52:1690–7.

62. Dong X, Wu P, Sun X, Li W, Wan H, Yu J, et al. Intra-tumour 18F-FDG uptake heterogeneity decreases the reliability on target volume definition with positron emission tomography/computed tomography imaging. J. Med. Imaging Radiat. Oncol. 2015;59:338–45.

63. Geets X, Lee JA, Bol A, Lonneux M, Gregoire V. A gradient-based method for segmenting FDG-PET images: methodology and validation. Eur J Nucl Med Mol Imaging. 2007;34:1427–38.

64. Nelson A, Brockway K, Nelson A, Piper J. PET Tumor Segmentation: Validation of a Gradient-based Method Using a NSCLC PET Phantom. 2009.

65. Hofheinz F, Langner J, Petr J, Beuthien-Baumann B, Steinbach J, Kotzerke J, et al. An automatic method for accurate volume delineation of heterogeneous tumors in PET. Med. Phys. 2013;40:082503.

66. Hatt M, Cheze le Rest C, Descourt P, Dekker A, De Ruysscher D, Oellers M, et al. Accurate automatic delineation of heterogeneous functional volumes in positron emission tomography for oncology applications. Int J Radiat Oncol Biol Phys. 2010;77:301–8.

67. Brooks FJ, Grigsby PW. Current measures of metabolic heterogeneity within cervical cancer do not predict disease outcome. Radiat. Oncol. Lond. Engl. 2011;6:69.

68. Groheux D, Majdoub M, Tixier F, Le Rest CC, Martineau A, Merlet P, et al. Do clinical, histological or immunohistochemical primary tumour characteristics translate into different (18)F-FDG PET/CT volumetric and heterogeneity features in stage II/III breast cancer? Eur. J. Nucl. Med. Mol. Imaging. 2015;42:1682–91.

69. Brooks FJ, Grigsby PW. FDG uptake heterogeneity in FIGO IIb cervical carcinoma does not predict pelvic lymph node involvement. Radiat. Oncol. Lond. Engl. 2013;8:294.

70. Kidd EA, Grigsby PW. Intratumoral metabolic heterogeneity of cervical cancer. Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res. 2008;14:5236–41.

71. Yang F, Thomas MA, Dehdashti F, Grigsby PW. Temporal analysis of intratumoral metabolic heterogeneity characterized by textural features in cervical cancer. Eur. J. Nucl. Med. Mol. Imaging. 2013;40:716–27.

72. Soussan M, Orlhac F, Boubaya M, Zelek L, Ziol M, Eder V, et al. Relationship between tumor heterogeneity measured on FDG-PET/CT and pathological prognostic factors in invasive breast cancer. PloS One. 2014;9:e94017.

73. Larson SM, Erdi Y, Akhurst T, Mazumdar M, Macapinlac HA, Finn RD, et al. Tumor treatment response based on visual and quantitative changes in global tumor glycolysis using PET-FDG imaging. The visual response score and the change in total lesion glycolysis. Clin Positron Imaging. 1999;2:159–71.

74. Yan J, Lim JC-S, Loi HY, Khor LK, Sinha AK, Quek ST, et al. Impact of image reconstruction settings on Texture Features in 18F-FDG PET. J. Nucl. Med. Off. Publ. Soc. Nucl. Med. 2015;

75. Hatt M, Tixier F, Cheze Le Rest C, Pradier O, Visvikis D. Robustness of intratumour ^{18}F-FDG PET uptake heterogeneity quantification for therapy response prediction in oesophageal carcinoma. Eur. J. Nucl. Med. Mol. Imaging. 2013;40:1662–71.

76. Brooks FJ, Grigsby PW. The effect of small tumor volumes on studies of intratumoral heterogeneity of tracer uptake. J Nucl Med. 2014;55:37–42.

77. Sullivan DC, Obuchowski NA, Kessler LG, Raunig DL, Gatsonis C, Huang EP, et al. Metrology Standards for Quantitative Imaging Biomarkers. Radiology. 2015;277:813–25.

78. QIBA FDG-PET/CT Technical Committee. FDG-PET/CT as an Imaging Biomarker Measuring Response to Cancer Therapy, Quantitative Imaging Biomarkers Alliance, Version 1.05. [Internet]. RSNA; 2013. Available from: RSNA.ORG/QIBA

79. Yip S, McCall K, Aristophanous M, Chen AB, Aerts HJWL, Berbeco R. Comparison of texture features derived from static and respiratory-gated PET images in non-small cell lung cancer. PloS One. 2014;9:e115510.

80. Oliver JA, Budzevich M, Zhang GG, Dilling TJ, Latifi K, Moros EG. Variability of Image Features Computed from Conventional and Respiratory-Gated PET/CT Images of Lung Cancer. Transl. Oncol. 2015;8:524–34.

81. Cheng N-M, Fang Y-HD, Tsan D-L, Hsu C-H, Yen T-C. Respiration-Averaged CT for Attenuation Correction of PET Images - Impact on PET Texture Features in Non-Small Cell Lung Cancer Patients. PloS One. 2016;11:e0150509.

82. Tixier F, Vriens D, Le Rest CC, Hatt M, Disselhorst JA, Oyen WJG, et al. Comparison of tumor uptake heterogeneity characterization between static and parametric 18F-FDG PET images in Non-Small Cell Lung Cancer. J Nucl Med. 2016;in press.

83. Nyflot MJ, Yang F, Byrd D, Bowen SR, Sandison GA, Kinahan PE. Quantitative radiomics: impact of stochastic effects on textural feature analysis implies the need for standards. J. Med. Imaging Bellingham Wash. 2015;2:041002.

84. Pyka T, Bundschuh RA, Andratschke N, Mayer B, Specht HM, Papp L, et al. Textural features in pre-treatment [F18]-FDG-PET/CT are correlated with risk of local recurrence and disease-specific survival in early stage NSCLC patients receiving primary stereotactic radiation therapy. Radiat. Oncol. Lond. Engl. 2015;10:100.

85. Cook GJR, O'Brien ME, Siddique M, Chicklore S, Loi HY, Sharma B, et al. Non-Small Cell Lung Cancer Treated with Erlotinib: Heterogeneity of (18)F-FDG Uptake at PET-Association with Treatment Response and Prognosis. Radiology. 2015;276:883–93.

86. Mu W, Chen Z, Liang Y, Shen W, Yang F, Dai R, et al. Staging of cervical cancer based on tumor heterogeneity characterized by texture features on (18)F-FDG PET images. Phys. Med. Biol. 2015;60:5123–39.

87. Oh JS, Kang BC, Roh J-L, Kim JS, Cho K-J, Lee S-W, et al. Intratumor Textural Heterogeneity on Pretreatment (18)F-FDG PET Images Predicts Response and Survival After Chemoradiotherapy for Hypopharyngeal Cancer. Ann. Surg. Oncol. 2015;22:2746–54.

88. Cheng N-M, Fang Y-HD, Chang JT-C, Huang C-G, Tsan D-L, Ng S-H, et al. Textural features of pretreatment 18F-FDG PET/CT images: prognostic significance in patients with advanced T-stage oropharyngeal squamous cell carcinoma. J. Nucl. Med. Off. Publ. Soc. Nucl. Med. 2013;54:1703–9.

89. Lovinfosse P, Janvary ZL, Coucke P, Jodogne S, Bernard C, Hatt M, et al. FDG PET/CT texture analysis for predicting the outcome of lung cancer treated by stereotactic body radiation therapy. Eur. J. Nucl. Med. Mol. Imaging. 2016;

90. Yip SSF, Coroller TP, Sanford NN, Mamon H, Aerts HJWL, Berbeco RI. Relationship between the Temporal Changes in Positron-Emission-Tomography-Imaging-Based Textural Features and Pathologic Response and Survival in Esophageal Cancer Patients. Front. Oncol. 2016;6:72.

91. Cheng N-M, Fang Y-HD, Lee L, Chang JT-C, Tsan D-L, Ng S-H, et al. Zone-size nonuniformity of 18F-FDG PET regional textural features predicts survival in patients with oropharyngeal cancer. Eur. J. Nucl. Med. Mol. Imaging. 2015;42:419–28.

92. Xu R, Kido S, Suga K, Hirano Y, Tachibana R, Muramatsu K, et al. Texture analysis on (18)F-FDG PET/CT images to differentiate malignant and benign bone and soft-tissue lesions. Ann. Nucl. Med. 2014;28:926–35.

93. Wu J, Aguilera T, Shultz D, Gudur M, Rubin DL, Loo BW, et al. Early-Stage Non-Small Cell Lung Cancer: Quantitative Imaging Characteristics of (18)F Fluorodeoxyglucose PET/CT Allow Prediction of Distant Metastasis. Radiology. 2016;151829.

94. Ypsilantis P-P, Siddique M, Sohn H-M, Davies A, Cook G, Goh V, et al. Predicting Response to Neoadjuvant Chemotherapy with PET Imaging Using Convolutional Neural Networks. PloS One. 2015;10:e0137036.

95. van Rossum PSN, Fried DV, Zhang L, Hofstetter WL, van Vulpen M, Meijer GJ, et al. The incremental value of subjective and quantitative assessment of 18F-FDG PET for the prediction of pathologic complete response to preoperative chemoradiotherapy in esophageal cancer. J. Nucl. Med. 2016;

96. Gao X, Chu C, Li Y, Lu P, Wang W, Liu W, et al. The method and efficacy of support vector machine classifiers based on texture features and multi-resolution histogram from (18)F-FDG PET-CT images for the evaluation of mediastinal lymph nodes in patients with lung cancer. Eur. J. Radiol. 2015;84:312–7.

97. Desseroit M-C, D. Visvikis, Tixier F, Majdoub M, Guillevin R, Perdrisot R, et al. Development of a nomogram combining clinical staging with 18F-FDG PET/CT image features in Non-Small Cell Lung Cancer stage I-III. Eur. J. Nucl. Med. Mol. Imaging. 2016;in press.

98. Fried DV, Mawlawi O, Zhang L, Fave X, Zhou S, Ibbott G, et al. Stage III Non-Small Cell Lung Cancer: Prognostic Value of FDG PET Quantitative Imaging Features Combined with Clinical Prognostic Factors. Radiology. 2015;142920.

99. Ohri N, Duan F, Snyder BS, Wei B, Machtay M, Alavi A, et al. Pretreatment 18FDG-PET Textural Features in Locally Advanced Non-Small Cell Lung Cancer: Secondary Analysis of ACRIN 6668/RTOG 0235. J. Nucl. Med. 2016;jnumed.115.166934.

100. Hyun SH, Kim HS, Choi SH, Choi DW, Lee JK, Lee KH, et al. Intratumoral heterogeneity of 18F-FDG uptake predicts survival in patients with pancreatic ductal adenocarcinoma. Eur. J. Nucl. Med. Mol. Imaging. 2016;1–8.

101. Lartizien C, Rogez M, Niaf E, Ricard F. Computer-aided staging of lymphoma patients with FDG PET/CT imaging based on textural information. IEEE J. Biomed. Health Inform. 2014;18:946–55.

102. Bang J-I, Ha S, Kang S-B, Lee K-W, Lee H-S, Kim J-S, et al. Prediction of neoadjuvant radiation chemotherapy response and survival using pretreatment [(18)F]FDG PET/CT scans in locally advanced rectal cancer. Eur. J. Nucl. Med. Mol. Imaging. 2015;

103. Vaidya M, Creach KM, Frye J, Dehdashti F, Bradley JD, El Naqa I. Combined PET/CT image characteristics for radiotherapy tumor response in lung cancer. Radiother. Oncol. J. Eur. Soc. Ther. Radiol. Oncol. 2012;102:239–45.

104. Antunes J, Viswanath S, Rusu M, Valls L, Hoimes C, Avril N, et al. Radiomics Analysis on FLT-PET/MRI for Characterization of Early Treatment Response in Renal Cell Carcinoma: A Proof-of-Concept Study. Transl. Oncol. 2016;9:155–62.

105. Wang H-M, Cheng N-M, Lee L-Y, Fang Y-HD, Chang JT-C, Tsan D-L, et al. Heterogeneity of (18) F-FDG PET combined with expression of EGFR may improve the prognostic stratification of advanced oropharyngeal carcinoma. Int. J. Cancer J. Int. Cancer. 2016;138:731–8.

106. Parmar C, Grossmann P, Rietveld D, Rietbergen MM, Lambin P, Aerts HJWL. Radiomic Machine-Learning Classifiers for Prognostic Biomarkers of Head and Neck Cancer. Front. Oncol. 2015;5:272.

107. Parmar C, Grossmann P, Bussink J, Lambin P, Aerts HJWL. Machine Learning methods for Quantitative Radiomic Biomarkers. Sci. Rep. 2015;5:13087.

108. Yoon HJ, Sohn I, Cho JH, Lee HY, Kim J-H, Choi Y-L, et al. Decoding Tumor Phenotypes for ALK, ROS1, and RET Fusions in Lung Adenocarcinoma Using a Radiomics Approach. Medicine (Baltimore). 2015;94:e1753.

109. Upadhaya T, Morvan Y, Stindel E, Le Reste P-J, Hatt M. A framework for multimodal imaging-based prognostic model building: Preliminary study on multimodal MRI in Glioblastoma Multiforme. IRBM [Internet]. [cited 2015 Oct 14]; Available from: http://www.sciencedirect.com/science/article/pii/S1959031815000962

110. Wang J, Kato F, Oyama-Manabe N, Li R, Cui Y, Tha KK, et al. Identifying Triple-Negative Breast Cancer Using Background Parenchymal Enhancement Heterogeneity on Dynamic Contrast-Enhanced MRI: A Pilot Radiomics Study. PloS One. 2015;10:e0143308.

111. Cameron A, Khalvati F, Haider M, Wong A. MAPS: A Quantitative Radiomics Approach for Prostate Cancer Detection. IEEE Trans. Biomed. Eng. 2015;

112. Khalvati F, Wong A, Haider MA. Automated prostate cancer detection via comprehensive multi-parametric magnetic resonance imaging texture feature models. BMC Med. Imaging. 2015;15:27.

113. Sharma RR, Marikkannu P. Hybrid RGSA and Support Vector Machine Framework for Three-Dimensional Magnetic Resonance Brain Tumor Classification. ScientificWorldJournal. 2015;2015:184350.

114. Mi H, Petitjean C, Dubray B, Vera P, Ruan S. Robust feature selection to predict tumor treatment outcome. Artif. Intell. Med. 2015;64:195–204.

115. Pandis N, Fedorowicz Z. The international EQUATOR network: enhancing the quality and transparency of health care research. J. Appl. Oral Sci. Rev. FOB. 2011;19.

Table 1. Properties of an ideal heterogeneity metric

Properties	Recommended methodology for evaluation
Repeatable	Compare metrics calculated on test-retest PET/CT images [51,57] using *e.g.* Bland-Altman method
Reproducible / robust	Compare metrics calculated through various analysis pipelines (with/without pre-processing such as denoising or partial volume effect correction, various segmentation approaches…) [57,75]
Least redundant with other TA metrics (and other variables).	Quantify and rank statistical correlation between features [59,60,75]. Use machine learning techniques to select features and combine them with other variables [107,114].
Offers value in regard of a given clinical endpoint	Quantify correlation with response to treatment, diagnosis, survival, differentiation of tumor types using robust machine learning techniques for classification, logistic regression and multivariate analysis, and learning / testing in separate cohorts (at least considering leave-one-out cross-validation) [28,107,114].

Table 2. Objectives

1.	Organize and develop a benchmark standard for TA metrics. This would include standardized physical and digital reference objects, open-source verified formulae and codes tested again reference objects, and expected values/results for comparison.
2.	Generate and circulate draft recommendations on image pre-processing, analysis and standardization, especially within the context of multicentric studies. Convene consensus groups to review, revise and ratify.
3.	Establish recommendations on methodological choices regarding the calculation of TA metrics and identify repeatable, reproducible and meaningful features (as well as their optimal calculation).
4.	Share publicly available cohorts of patients with PET/CT images and associated clinical data, along with clinical endpoint (survival, response to therapy, tumour type classification, etc.) so that research groups can test/evaluate their workflow
5.	Support larger prospective multicentric studies and the use of robust statistical analysis by exploiting the methods from the field of machine learning.
6.	Adopt standards for publishing methods and results such as those promoted by the EQUATOR (Enhancing the QUAlity and Transparency Of health Research) network [115].
7.	Advocate for improved peer-review, insisting on at least one “statistical reviewer” with knowledge of machine learning methodologies.

Figure legends

Figure 1: Workflow involved in the calculation and selection of texture analysis from a reconstructed PET image.

Figure 2: Distributions with respect to (A) MATV and (B) SUV_{max} of four TA features (correlation and $entropy_{GLCM}$ from GLCM, complexity from NGTDM and zone size percentage from GLZSM) calculated after either quantization into a set number of bins (here 64) or into bins of fixed width (here 0.5 SUV). Note that correlation is not impacted, compared to the three other metrics. Also, note the inverted correlation with MATV and SUV_{max}, when changing the quantization approach.

Figure 3: Distribution of heterogeneity as measured with $entropy_{GLCM}$ with respect to MATV for 555 lesions in five tumour types, according to four different configurations: with a quantization of either 64 or 128 grey-levels (uniformly distributed) and using either 1 single co-occurrence matrix (without averaging) or 13 matrices followed by averaging.

Figure 4: Illustration of trade-offs in segmentation results using (A) contour-based approach (PETedge from MIMVista software, white external contour) or (B) a clustering-based approach (the FLAB algorithm, blue contours). In (A) the light grey contour inside the tumour corresponds to a fixed threshold. With (B) the various areas with different uptake levels are automatically determined and may be included or excluded from the heterogeneity analysis at the cost of a higher complexity.

Figure 5: illustration of the relationship between a textural feature and the corresponding MATV in 116 NSCLC patients (A linear scale and B log scale) and the resulting quantification of the correlation according to Pearson coefficients and different rank coefficients (Spearman and Kendall).

Figure 6: distributions of heterogeneity ($entropy_{GLCM}$) with respect to MATV for a set of 25 tumours reconstructed with either 4×4×4 mm^3 (A and B) or 2×2×2 mm^3 (C) voxels. Note that the 4×4×4 mm^3 voxels image was upsampled to 2×2×2 mm^3 using either nearest neighbors (A) or B-spline (B) interpolation.

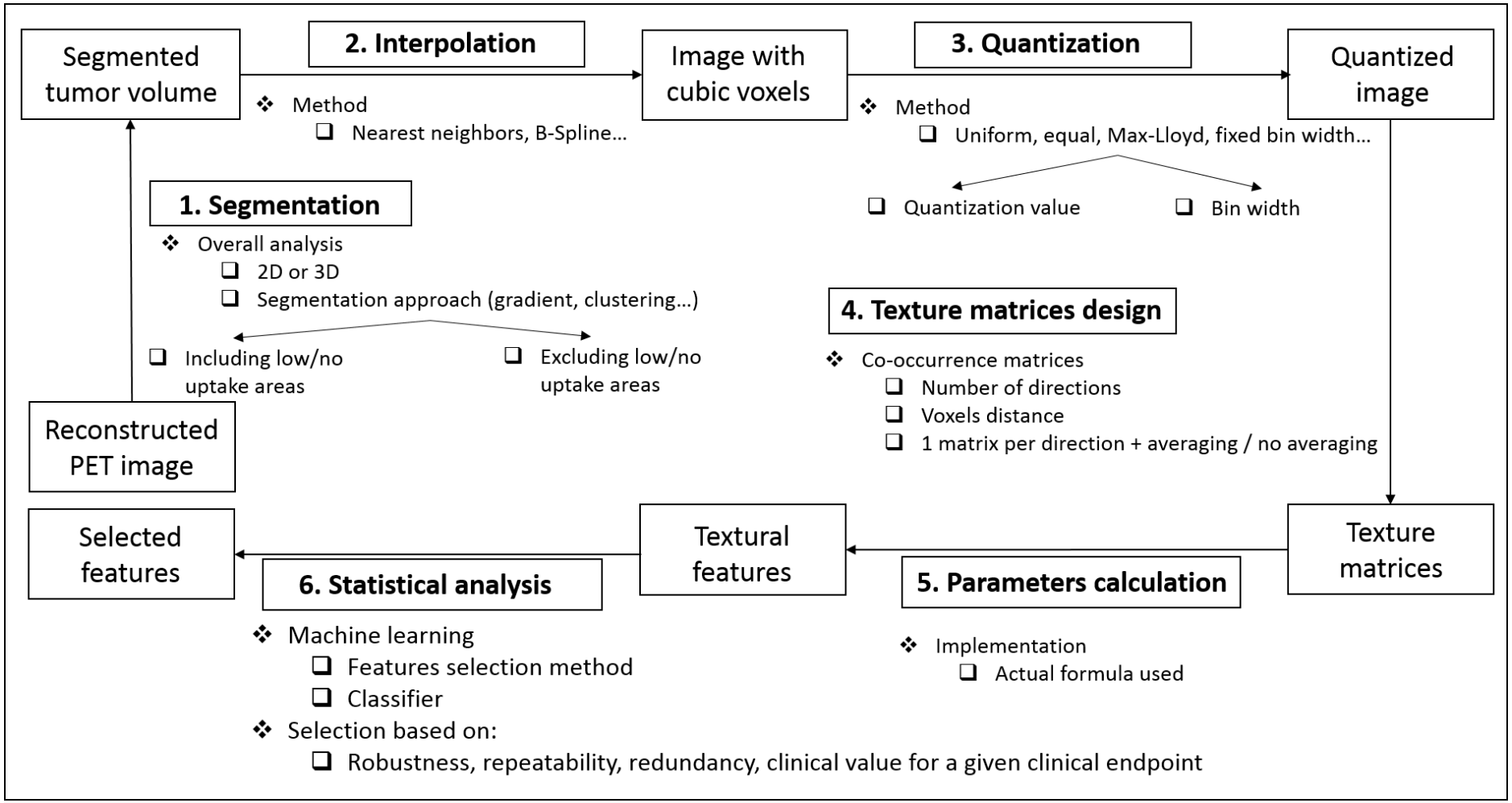
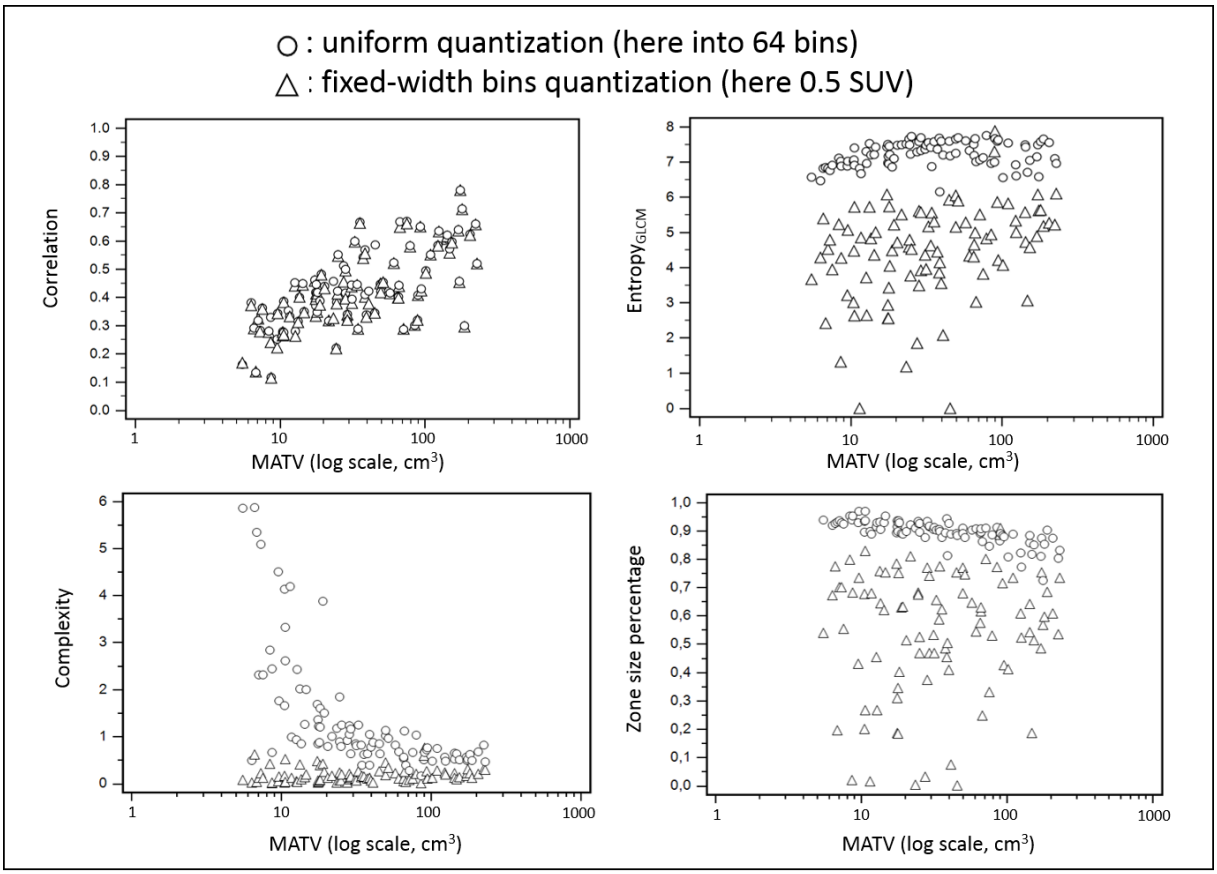

Segmented tumor volume
2. Interpolation
Method
Nearest neighbors, B-Spline…
Image with cubic voxels
3. Quantization
Method
Uniform, equal, Max-Lloyd, fixed bin width…
Quantization value
Bin width
Quantized image
1. Segmentation
Overall analysis
2D or 3D
Segmentation approach (gradient, clustering…)
Including low/no uptake areas
Excluding low/no uptake areas
4. Texture matrices design
Co-occurrence matrices
Number of directions
Voxels distance
1 matrix per direction + averaging / no averaging
Reconstructed PET image
Selected features
6. Statistical analysis
Machine learning
Features selection method
Classifier
Selection based on:
Robustness, repeatability, redundancy, clinical value for a given clinical endpoint
Textural features
5. Parameters calculation
Implementation
Actual formula used
Texture matrices

Figure 1

A


○ : uniform quantization (here into 64 bins)
△ : fixed-width bins quantization (here 0.5 SUV)
Correlation
EntropyGLCM
Complexity
Zone size percentage
MATV (log scale, cm3)
MATV (log scale, cm3)
MATV (log scale, cm3)
MATV (log scale, cm3)

B

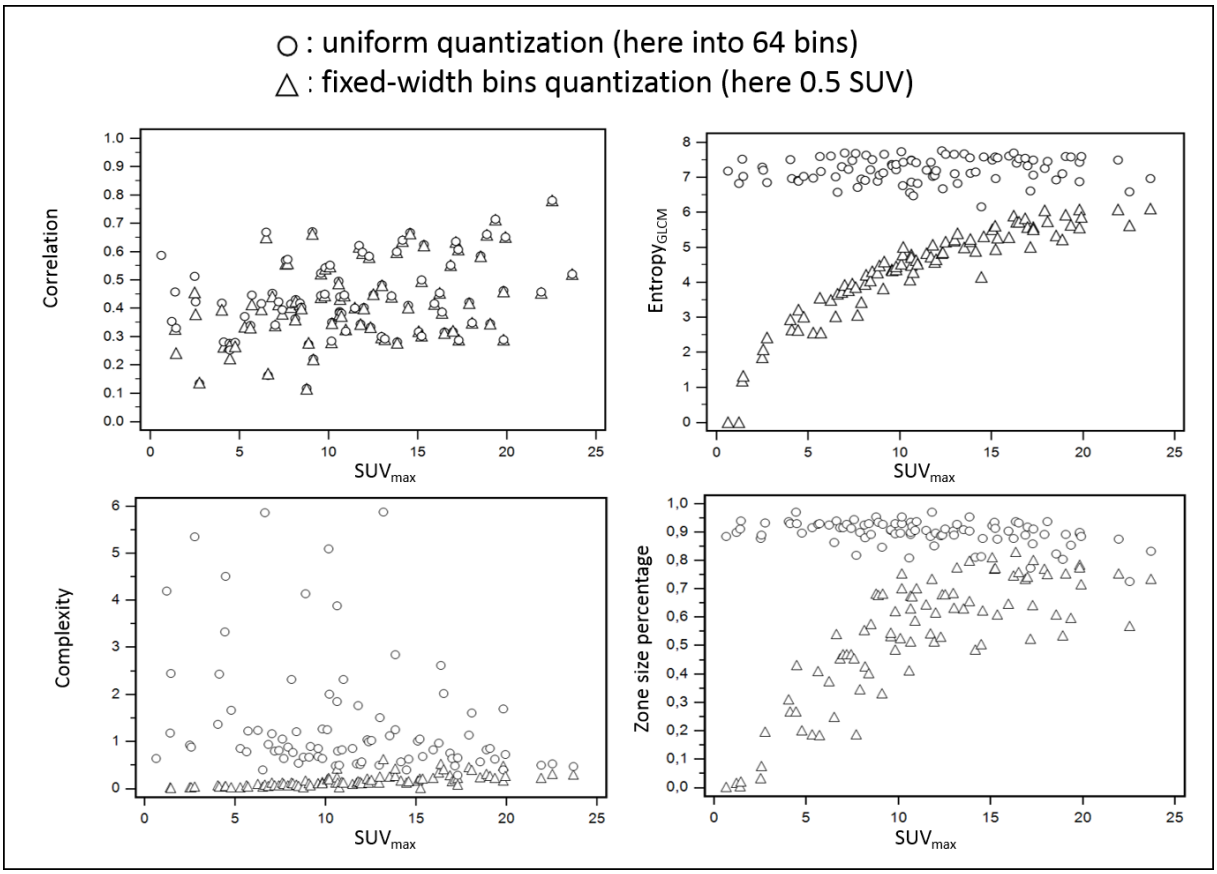

○ : uniform quantization (here into 64 bins)
△ : fixed-width bins quantization (here 0.5 SUV)
Correlation
Entropy$_{GLCM}$
Complexity
Zone size percentage
SUV$_{max}$
SUV$_{max}$
SUV$_{max}$
SUV$_{max}$

Figure 2

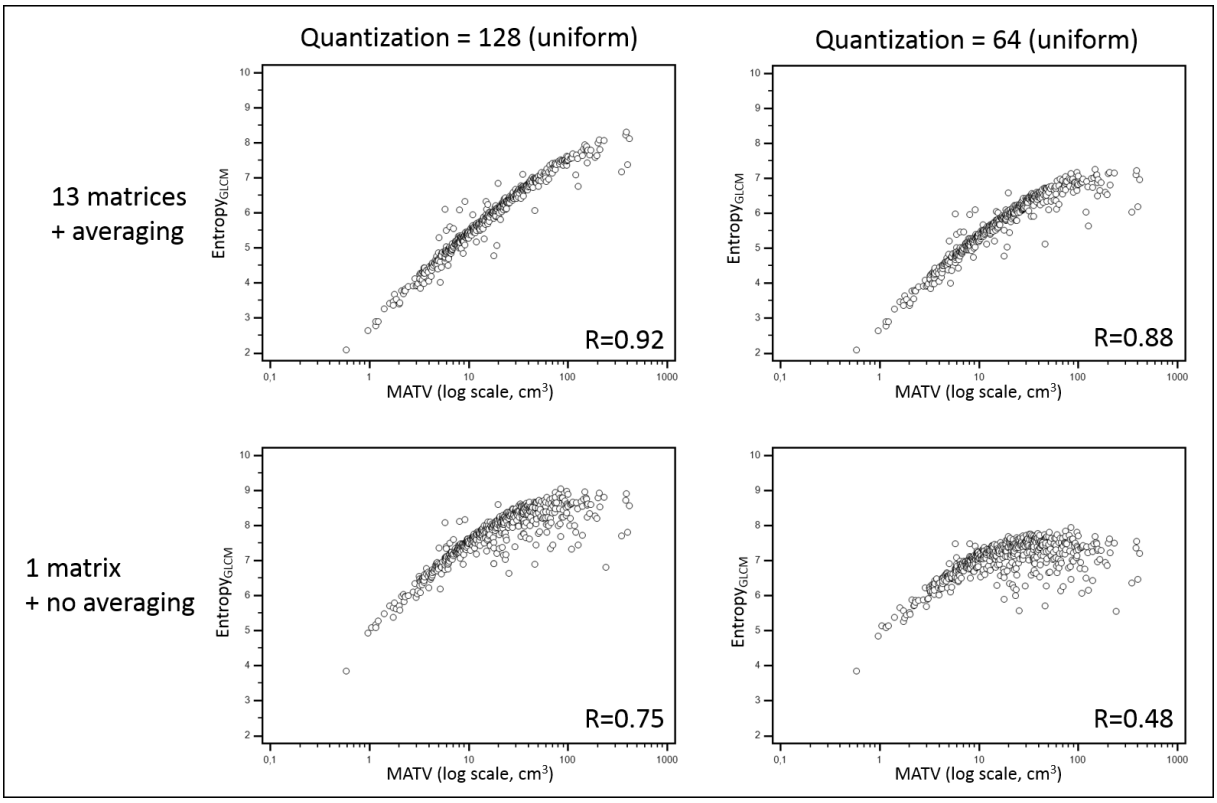

Quantization = 128 (uniform)
Quantization = 64 (uniform)
13 matrices
+ averaging
1 matrix
+ no averaging
Entropy$_{GLCM}$
MATV (log scale, cm^3)
R=0.92
R=0.88
R=0.75
R=0.48

Figure 3

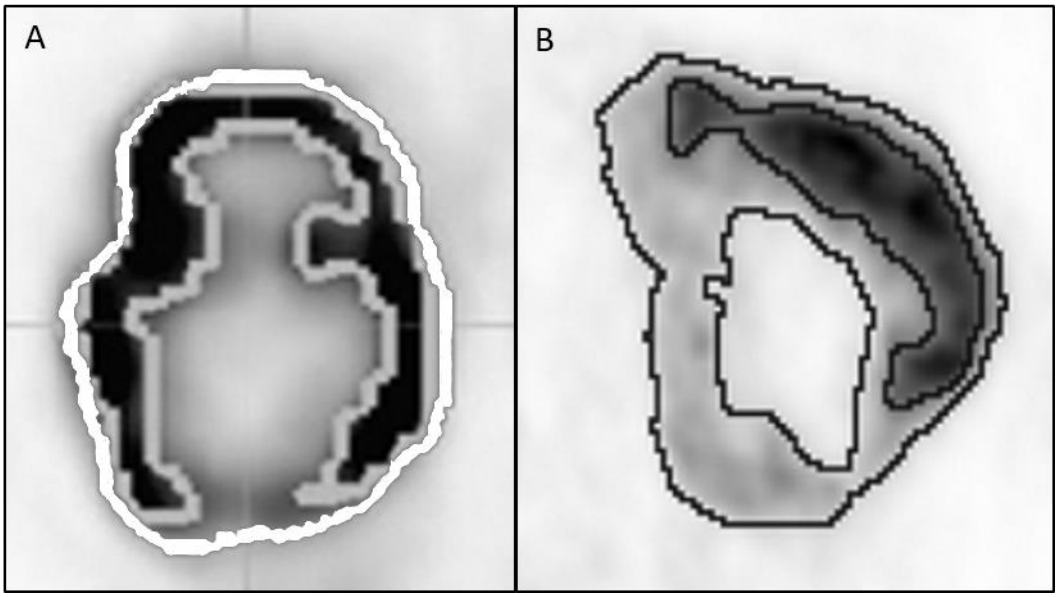


Figure 4

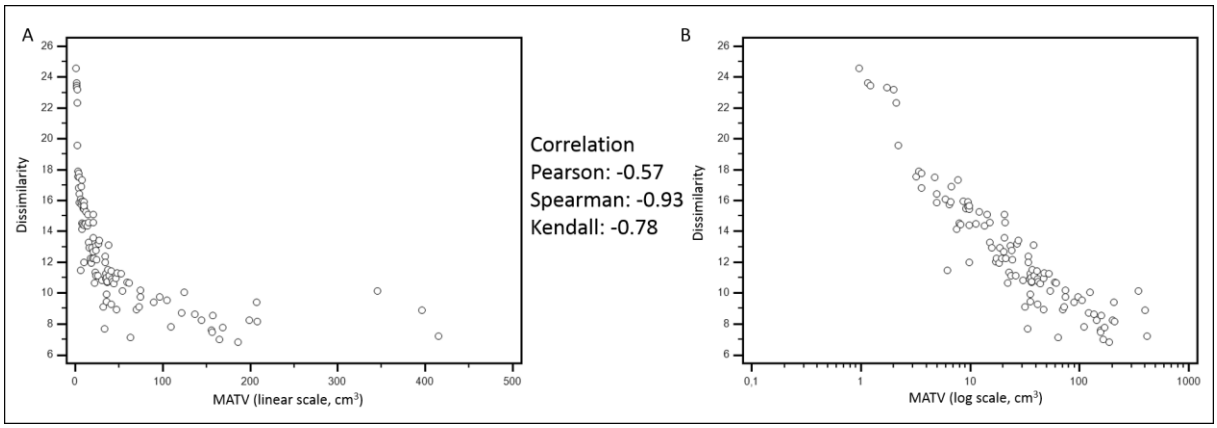

A
Dissimilarity
MATV (linear scale, cm3)
Correlation
Pearson: -0.57
Spearman: -0.93
Kendall: -0.78
B
Dissimilarity
MATV (log scale, cm3)

Figure 5

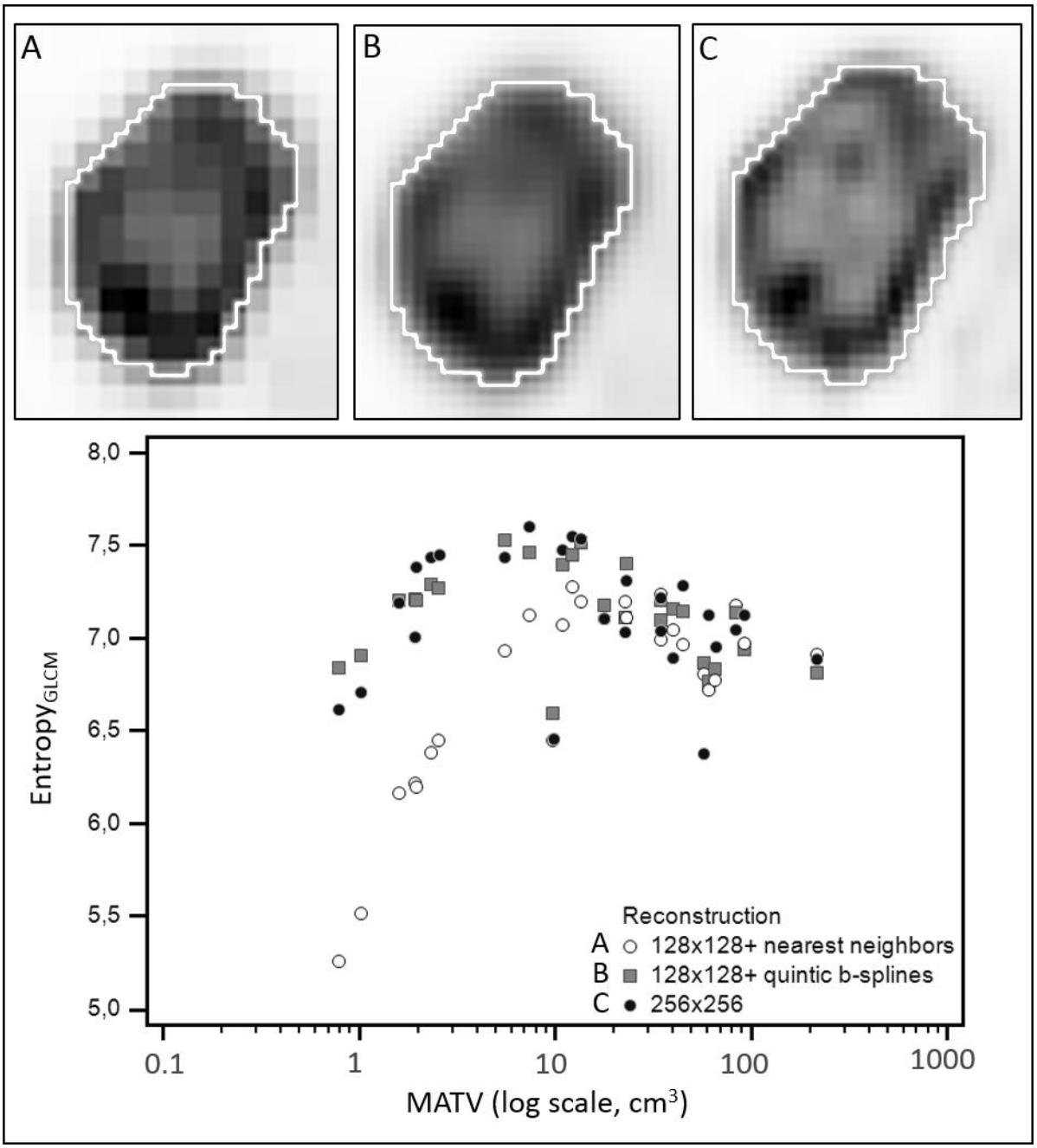

A
B
C
8,0
7,5
7,0
6,5
6,0
5,5
5,0
Entropy$_{GLCM}$
0.1
1
10
100
1000
MATV (log scale, cm^3)
Reconstruction
A 128x128+ nearest neighbors
B 128x128+ quintic b-splines
C 256x256

Figure 6